Spatial and Spatiotemporal Point Process Modelling in Epidemiology

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

at Massey University, Manawatu
New Zealand

Tilman Marcus Davies

December, 2011

Abstract

Geographical epidemiology encapsulates those problems that aim to understand space and/or space time trends in the disease(s) of interest, a goal clearly important in both public health and economic contexts. The increasing availability of point-location, coordinate data for studies in geographical epidemiology calls for a greater scrutiny of statistical point process theory with respect to these applications. Though the statistical analysis of planar point patterns is now theoretically well-established, there remain many aspects which warrant further research, especially from a practical perspective. The need becomes even greater when we consider the relative youth of methods for the analysis of spatiotemporal observations, where non-trivial variation and even dependence throughout the space-time continuum can exist.

This work aims to address these issues by careful review, theoretical refinement, empirical testing, and real-world analyses of certain statistical tools used in point process problems in geographical epidemiology. We scrutinise the kernel-smoothed density-ratio estimator of the so called *relative risk function*, a particularly flexible approach given the anticipated spatial heterogeneity of the observations over a given geographical region. This discussion introduces the adaptive (i.e. variable bandwidth) risk function, as well as novel asymptotic methods for computation of tolerance contours designed to identify sub-regions of statistically significant fluctuations in risk.

More sophisticated statistical methodology is warranted in certain situations, where it may be assumed that both 'global' heterogeneity and 'local' correlation drives the space and/or space-time disease dispersion. A comprehensive review of the stochastic *log-Gaussian Cox process*, in both purely spatial and spatiotemporal contexts, is conducted. A suite of novel numerical experiments investigate the performance of convenient, yet *ad hoc*, minimum contrast parameter estimation techniques for the dependence structure of the latent Gaussian process.

The computer code arising from the review and refinement of the above methodologies was instrumental in the release of two separate software packages. These are available in the R environment, and also showcased here. A number of additional collaborations with applied researchers around the world serve to further highlight the contributions made throughout the course of this research project and the importance of sound statistical methods in geographical epidemiology.

Acknowledgements

To my amazing family, both across the Tasman and up in Germany, for their unwavering and continuous support;

To my wonderful cat, Sigma, for always being happy to see me and for maintaining a consistent supply of fluff and semi-deceased birds;

To the statistics academics across Europe for their interest in my work and generous hospitality during my trip in early 2011;

To my external collaborators of various peer-reviewed works that took an interest in the contents of this thesis while still in production; this academic experience at such an early stage of my career was indeed invaluable, and of course extremely rewarding. Special thanks are reserved for M. Stevenson, EpiCentre, Massey University (New Zealand), B. Taylor, B. Rowlingson and P. Diggle, Division of Medicine, Lancaster University (U.K.), and Z. Zhang, Department of Epidemiology, Fudan University (People's Republic of China);

To my local co-supervisors, Associate Professor Geoff Jones and Doctor Jonathan C. Marshall, a massive thanks for never failing to provide that extra support and additional points of view absolutely invaluable to understanding and implementing novel and/or complex material;

And most importantly to my primary supervisor, Professor Martin L. Hazelton, without whom this work quite simply would not exist. Responsible not just for this project's supervision, but for inspiring me in undergraduate years to further pursue statistics in the first place, he ensured that the positive times far outweighed the negative. True intelligence is not just knowledge, but the ability to communicate, and teach, said knowledge. Martin is, without a doubt, the most intelligent academic I've ever met.

Contents

List of Figures

List of Tables

List of Publications

Accredited publications arising during the course of this research.

1. Taylor BM, **Davies TM**, Rowlingson BS, Diggle PJ (2012) **lgcp** - An R package for inference with spatial and spatiotemporal log-Gaussian Cox processes, *Journal of Statistical Software* [in press].

2. Zhang ZJ, Chen DM, Chen Y, **Davies TM**, Vaillancourt JP, Liu WB (2012) Risk signals of an influenza pandemic caused by highly pathogenic avian influenza subtype H5N1: Spatio-temporal perspectives, *The Veterinary Journal* **192** 3 417-421.

3. Zhang ZJ, **Davies TM**, Jiang QW (2012) Identification of high-risk regions for schistosomiasis in the Guichi region of China: an adaptive kernel density estimation-based approach, *Submitted for publication – Geospatial Health*.

4. **Davies TM**, Hazelton ML, Marshall JC (2011) **sparr**: Analyzing spatial relative risk using fixed and adaptive kernel density estimation in R, *Journal of Statistical Software* **39** 1 1-14.

5. Sanson RL, Harvey N, Garner MG, Stevenson MA, **Davies TM**, Hazelton ML, O'Connor J, Dubé C, Forde-Folle KN, Owen K (2011) Foot-and-mouth disease model verification and 'relative validation' through a formal model comparison, *Revue Scientifique et Technique-Office International des Epizooties* **30** 2 527-540.

6. **Davies TM**, Hazelton ML (2010) Adaptive kernel estimation of spatial relative risk, *Statistics in Medicine* **29** 23 2423-2437.

7. Hazelton ML, **Davies TM**, (2009) Inference based on kernel estimates of the relative risk function in geographical epidemiology, *Biometrical Journal* **51** 1 98-109 (**Pre-Ph.D.**).

List of Acronyms

Definitions of acronyms and abbreviations, and the section of their first occurrence in the thesis.

2D	Two-dimensional	1.2
3D	Three-dimensional	4.1.2
ADAPT	Adaptive-bandwidth relative risk function	2.5
AEGISS	Ascertainment and Enhancement of Gastrointestinal Infection Surveillance and Statistics	3.4.2
AMISE	Asymptotic mean integrated squared error	2.3.1
ARE	Absolute relative error	3.5.1
ASY	Asymptotic	2.4
BCV	Biased cross-validation	2.3.1
CRAN	Comprehensive R Archive Network	1.3
CSR	Complete spatial randomness	1.2.1
CV	Cross-validation	2.3.1
DFT	Discrete Fourier transform	3.3
DMCS	Deterministic Minimum Contrast Scenario	3.5.2
ETAS	Epidemic type aftershock sequence	5.4
EU	European Union	5.3.2
EW	Epidemic wave	5.2.2
FAO	Food and Agriculture Organisation of the United Nations	5.2.2
FFT	Fast-Fourier transform	3.3
FIXCV	Fixed-bandwidth relative risk function (CV bandwidth)	2.5
FIXOS	Fixed-bandwidth relative risk function (OS bandwidth)	2.5
FMD	Foot-and-mouth disease	5.3.1
GCMT	Global Centroid Moment Tensor	5.4
GPS	Global positioning system	5.1.2
iid	Independent and identically distributed	2.1.1
IQR	Interquartile range	2.3.1
ISE	Integrated squared error	2.3.1
KDE	Kernel density estimation	2.1
LGCP	Log-Gaussian Cox process	3.1
LOWESS	Locally weighted scatterplot smoothing	3.4.2
LSCV	Least-squares cross-validation	2.3.1
MALA	Metropolis-adjusted Langevin algorithm	3.3
MC	Monte-Carlo	2.4
MCMC	Markov-chain Monte Carlo	3.3.2
MCS	Minimum Contrast Scenario	3.5.1
MH	Metropolis-Hastings	3.3.2
MISE	Mean integrated squared error	2.3.1
NAADSM	North American Animal Disease Spread Model	5.3.1
netCDF	Network common data form	4.2.2
OIE	World Organisation for Animal Health	5.2.2
OS	Oversmoothing (maximal smoothing principle)	2.3.1
PBC	Primary biliary cirrhosis	2.6.2
PCF	Pair correlation function	1.2.3
PDF	Probability density function	2.1.1
RED	Relative Euclidian distance	3.5.2
SMR	Standardised mortality ratio	1.1
SNCP	Shot-noise Cox process	3.1.4
STS	SpatioTemporal Scenario	3.5.3
WISE	Weighted integrated squared error	2.5
WMISE	Weighted mean integrated squared error	2.3.1